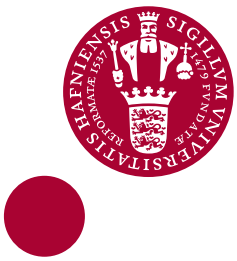


Prediction of pKa values for druglike molecules using semiempirical quantum chemical methods

Jensen, Jan Halborg; Swain, Christopher J; Olsen, Lars

Published in:
Journal of Physical Chemistry Part A: Molecules, Spectroscopy, Kinetics, Environment and General Theory

DOI:
10.1021/acs.jpca.6b10990

Publication date:
2017

Document version
Early version, also known as pre-print

Prediction of pKa Values for Drug-Like Molecules Using Semiempirical Quantum Chemical Methods

Jan H. Jensen,*,† Christopher J. Swain,‡ and Lars Olsen¶

†*Department of Chemistry, University of Copenhagen, Copenhagen, Denmark*

‡*Cambridge MedChem Consulting, Cambridge, UK*

¶*Section of Biostructural Research, Department of Drug Design and Pharmacology, University of Copenhagen, Copenhagen, Denmark*

E-mail: jhjensen@chem.ku.dk;Twitter:@janhjensen

Abstract

Rapid yet accurate pKa prediction for drug-like molecules is a key challenge in computational chemistry. This study uses PM6-DH+/COSMO, PM6/COSMO, PM7/COSMO, PM3/COSMO, AM1/COSMO, PM3/SMD, AM1/SMD, and DFTB3/SMD to predict the pKa values of 53 amine groups in 48 drug-like compounds. The approach uses an isodesmic reaction where the pKa value is computed relative to a chemically related reference compound for which the pKa value has been measured experimentally or estimated using a standard empirical approach. The AM1- and PM3-based methods perform best with RMSE values of 1.4 - 1.6 pH units that have uncertainties of ±0.2-0.3 pH units, which make them statistically equivalent. However, for all but PM3/SMD and AM1/SMD the RMSEs are dominated by a single outlier, cefadroxil, caused by proton transfer in the zwitterionic protonation state. If this outlier is removed, the RMSE values for PM3/COSMO and AM1/COSMO drop to 1.0 ± 0.2 and 1.1 ± 0.3, while PM3/SMD and AM1/SMD remain at 1.5 ± 0.3 and 1.6 ± 0.3/0.4 pH units, making the COSMO-based predictions statistically better than the SMD-based predictions. So for pKa calculations where a zwitterionic state is not involved or proton transfer in a zwitterionic state is not observed then PM3/COSMO or AM1/COSMO is the best pKa prediction method, otherwise PM3/SMD or AM1/SMD should be used. Thus, fast and relatively accurate pKa prediction for 100-1000s of drug-like amines is feasible with the current setup and relatively modest computational resources.

Introduction

One of the central practical challenges to be met when performing calculations on many organic molecules in aqueous solution is selecting the correct protonation state at a given pH. There are several empirical pKa predictors such as ACD pKa DB (ACDLabs, Toronto, Canada), ChemAxon (ChemAxon, Budapest, Hungary), and Epik (Schrödinger, New York, USA) that rely on large databases of experimental pKa values that are adjusted using empirical substituent-specific rules. As with any empirical approach the accuracy of these methods

correlate with the similarity of the target molecule to molecules in the database. For example, Settimo et al.[1] have recently shown that the empirical methods can fail for some amines, which represent a large fraction of drugs currently on the market or in development. This problem could make these methods difficult to apply to computational exploration of chemical space[2–4] where molecules with completely novel chemical substructures are likely to be encountered.

One possible solution to this problem is electronic structure (QM)-based pKa prediction methods (see Ho[5] for a review) which in principle requires no empirical input. In practice, when applied to larger molecules[6–8], some degree of empiricism is usually introduced to increase the accuracy of the predictions but these parameters tend to be much more transferable because of the underlying QM-model. However, these QM-based methods are computationally quite demanding and cannot be routinely applied to the very large sets of molecules typically encountered in high throughput screening.

Semiempirical QM methods such as PM6[9] and DFTB3[10] are orders of magnitude faster than QM methods but retain a flexible and, in principle, more transferable QM description of the molecules. One of us recently co-authored a proof-of-concept study[11] demonstrating that semiemprical QM methods can be used together with isodesmic reactions to predict pKa values of small model systems with accuracies similar to QM methods for many functional groups. However, amines proved the most difficult due to the diverse chemical environment of the ionizable nitrogen atoms. We hypothesized that the solution to this problem is a more diverse set of reference molecules and in this study we demonstrate the validity of this hypothesis for a set of 53 amine groups in 48 drug-like compounds. In addition we test more semiempirical methods than in the previous study.

Computational Methodology

The pKa values are computed by

$$\mathrm{pK_a} = \mathrm{pK_a^{ref}} + \frac{\Delta G^\circ}{RT\ln(10)} \tag{1}$$

where ΔG° denotes the change in standard free energy for the isodesmic reaction

$$\mathrm{BH^+ + B_{ref} \rightleftharpoons B + B_{ref}H^+} \tag{2}$$

where the standard free energy of molecule X is computed as the sum of the semiempirical heat of formation, or the electronic energy in case of DFTB3, and the solvation free energy

$$G^\circ(X) = \Delta H_f^\circ(X) + \Delta G_{solv}^\circ(X) \tag{3}$$

All energy terms are computed using solution phase geometries unless noted otherwise. $\Delta H_f^\circ(X)$ is computed using either PM6-DH+[12], PM6[13], PM7[14], PM3[15], AM1[16], or DFTB3[10] (where the electronic energy is used instead of the heat of formation), while $\Delta G_{solv}^\circ(X)$ is computed using either the SMD[17] or COSMO[18] solvation method. The SMD calculations are performed with the GAMESS program[19], the latter using the semiempirical PCM interface developed by Steinmann et al.[20] and the DFTB/PCM interface developed by Nishimoto[21] and using version 3ob-3-1 of the 3OB parameter set[10,22–24], while the COSMO calculations are performed using MOPAC2016. A maximum of 200 optimization cycles are used for solution phase optimizations and a gradient convergence criterion (OPTTOL) of 5×10^{-4} au and delocalized internal coordinates[25] are used for GAMESS-based optimization.

This study considers 53 amine groups in 48 drug-like molecules with experimentally measured amine pKa values taken from Table 3 of the study by Eckert and Klamt[6]. Some of the smaller molecules in that table, such as 2-methylbenzylamine, were removed since they

would differ very little from the corresponding reference molecules. The reference molecules are chosen to match the chemical environment of the nitrogen within a two-bond radius as much as practically possible and including the ring-size if the nitrogen is situated in a ring. For example, the tertiary amine group in thenyldiamine (Figure 1) has two methyl groups and a longer aliphatic chain so the reference molecule is dimethylethylamine, rather than triethylamine used in our previous study. This choice is motivated by our previous observation[11] that, for example, the predicted value of dimethylamine has a relatively large error when computed using a diethylamine reference. Similarly, the reference compound for the aromatic nitrogen group in thenyldiamine is 2-aminopyridine, rather than pyridine, to reflect the fact that the nitrogen is bonded to an aromatic carbon which is bonded to another aromatic carbon and another nitrogen. In a few cases somewhat larger reference compounds are chosen if they reflect common structural motifs such as the guanine group in acyclovir or the $^{-}$OOC-CH(R)-NH_3^+ zwitterionic motif in phenylalanine and tryptophan. This approach resulted in 26 different reference molecules (Table S1) that reflect typical functional groups found in drug-like molecules. Most of the reference pKa values are computed using the ACE JChem pKa predictor[26] while the rest are experimental values. The only molecule where it proved difficult to apply this general approach to identifying a suitable reference molecule is the imine nitrogen in clozapine (Figure 1) where the nitrogen is bonded to a phenyl group on one side an a tertiary sp^2 carbon that in turn is hydrogen bonded to a nitrogen and a phenyl group. The reference compound that would result from applying the rules outlined above (N-phenylbenzamidine, Figure 1) was considered "too specific" for clozapine. Instead we searched the already chosen set of 26 reference molecules the molecule with the largest sub-structure match, which turn out to N-phenylethanimidamide, that was originally chosen as a reference for phenacaine.

Many of the molecules contain more than one ionizable group. Only the pKa values of the amine indicated in Eckert and Klamt's Table 3 are computed and the protonation

states are prepared according to standard pKa values. For example, for phenylalanine the carboxyl group is deprotonated because the "standard" pKa values of a carboxyl group (e.g. in acetic acid) is lower than the standard pKa values of a primary amine (e.g. ethylamine). Notice that the cyanoguanidine group in cimetidine has a pKa value of about 0[27,28] and is therefore deprotonated when the imidazole group titrates. Eckert and Klamt characterised the histamine pKa value of 9.7 as an amidine pKa and the thenyldiamine pKa as a pyridine pKa. This is corrected to a primary amine[29] and tertiary amine, respectively. For thenyldiamine pKa values of 3.7 and 8.9 have been measured potentiometrically[30] and cannot be assigned to a particular nitrogen experimentally. But based on standard pKa values it it is likely that the higher pKa value corresponds to the amino group. For example, the ACE JChem pKa predictor predicts values of 5.6 and 8.8 for the pyridine and amine groups, respectively. This hypothesis is further corroborated by the fact that introducing an additional N atom to the pyridine ring in neohetramine (Figure 1) only significantly affects the lower pKa value[30]. The experimental pKa values of morphine and niacin are changed to 8.2[31] and 4.2[32], respectively, while the remaining experimental pKa values are taken from Eckert and Klamt[6]. When several tautomers are possible all are considered. The protonation and tautomer states considered can be found in supplementary materials. RDKit[33] is used to generate 20 starting geometries for each protonation state and the lowest free energy structure for each protonation state is used for the pKa calculations.

The Epik[34,35] calculations where performed with version 2016-4 of the Epik program using coordinates generated from SMILES strings using LigPrep version 2016-4. Default settings were applied except that the initial ionization state was not changed. Solvent was selected as water and the pH range as 7.0 ± 2.0. The ChemAxon calculations are performed using the command line tool cxcalc version 15.12.14.0. The ACD predictions are taken from the ChEMBL20 database with except the pKa value of nikethamide, which is taken from ChEMBL19. Versions 20-22 lists a pKa value of 10.1 for nikethamide, which is considerably

higher than the experimental value of 3.5. Version 19 lists a pKa value of 4.01, which is in better agreement with experiment so this value was used. The ChEMBL pKa data is computed using ACDlabs software v12.01

Results and Discussion

PM3- and AM1-based methods

Table 1 lists the predicted pKa values, Figure 2 shows a plot of the errors, and Table 2 lists the root-mean-square-error (RMSE) and maximum absolute error for each method. The AM1- and PM3-based methods perform best with RMSE values of 1.4 - 1.6 pH units that have uncertainties in the 0.2-0.3 pH unit range, which make them statistically equivalent[36–38] (see SI for more information). The null model $pK_a \approx pK_a^{ref}$ has an RMSE of 1.8 ± 0.3/0.4. However, because of the high correlation between the null model and the PM3 and AM1 methods (e.g. $r = 0.78$ vs PM3/COSMO) the composite errors are relatively small (e.g. 0.2 pH units vs PM3/COSMO) making the lower RMSE observed for AM1 and PM3 statistically significant. The rest of the methods (PM6-DH+, PM6, PM7, and DFTB3) perform worse than AM1 and PM3 and are discussed further below.

The negative outlier seen for the COSMO-based methods (Figure 2) is cefadoxil (Figure 1) and is due to proton transfer in the zwitterionic protonation state. For the three other zwitterions among the molecules, niacin, phenylalanine, and tryptophan, no proton transfer is observed and the error in the predicted pKa values are relatively small. Proton transfer in zwitterions is also a common problem for DFT/continuum calculations, for example for glycine[39–41], and is due to deficiencies in the continuum solvent method, not the electronic structure method. The good performance observed for PM3/SMD is thus due to fortuitous cancellation of error. Cefadroxil is also the negative outlier for DFTB3/SMD although the proton does not transfer.

Table 1: Experimental, reference (cf. Eq 1), and predicted pKa values. "COS" stands for COSMO. "+" and "-" refers to the charge of the conjugate base.

Molecule	Exp	Ref pKa	PM6-DH+ COS	PM6 COS	PM7 COS	PM3 COS	AM1 COS	PM3 SMD	AM1 SMD	DFTB3 SMD
Acebutolol	9.5	10.6	8.2	6.8	10.4	9.4	9.2	6.8	8.8	10.5
Acyclovir	2.2	2.6	0.1	0.3	0.3	1.5	1.0	0.9	1.2	0.9
Alphaprodine	8.7	10.1	6.5	6.5	9.0	8.5	8.6	7.9	8.0	6.2
Alprenolol	9.6	10.6	6.8	6.8	9.1	8.1	8.8	7.4	7.9	8.6
Atenolol	9.6	10.6	8.4	7.2	9.9	9.0	9.4	7.8	8.2	8.1
Benzocaine	2.5	4.6	0.2	0.3	1.3	2.6	1.8	2.6	1.8	0.8
Betahistine	10.0	10.5	9.3	8.7	10.2	8.6	9.5	8.6	8.9	10.6
Betahistine+	3.9	5.8	2.9	3.8	3.1	4.9	3.3	4.5	2.8	-0.2
Cefadroxil-	7.0	9.5	-2.8	1.2	1.0	0.1	1.6	7.6	4.6	-3.0
Chloroquine	10.6	10.2	9.6	9.1	11.4	9.9	9.9	8.6	9.2	8.3
Cimetidine0	6.8	6.8	6.1	6.1	4.1	5.1	5.5	4.2	5.2	5.3
Clomipramine	9.4	10.2	11.2	10.0	13.6	9.7	9.2	9.4	8.5	8.4
Clotrimazole	5.8	6.6	5.1	5.0	7.3	4.7	5.2	4.0	4.0	4.3
Clozapine	7.5	10.0	5.9	5.7	7.2	8.2	7.8	7.1	7.1	6.2
Clozapine+	3.9	10.3	5.4	5.6	6.8	2.7	2.5	1.4	1.2	6.6
Codeine	8.1	10.1	6.0	5.7	8.0	7.6	6.2	6.1	4.7	5.4
Desipramine	10.3	10.5	10.8	10.9	11.9	9.9	9.5	9.8	8.9	9.4
Guanethidine	11.4	12.8	14.3	12.4	13.2	13.2	14.0	12.3	12.9	16.2
Histamine	9.7	10.6	9.5	9.5	9.7	8.9	9.9	9.0	9.4	12.9
Hydroquinine	9.1	10.5	7.0	6.4	10.0	8.9	8.9	6.9	8.4	9.7
Hydroquinine+	4.1	4.5	2.7	2.1	3.8	1.8	2.2	3.1	1.7	3.1
Imipramine	9.6	10.2	9.9	8.9	12.0	9.6	9.3	9.3	8.4	8.6
Labetalol	7.3	10.6	7.0	6.4	9.4	7.5	9.9	7.2	8.4	8.0
Lidocaine	7.9	10.2	3.7	4.2	5.7	5.4	5.7	5.3	5.3	5.9
Maprotiline	10.3	10.5	10.2	10.4	11.7	10.9	10.2	10.5	9.6	10.4
Mechlorethamine	6.4	10.0	4.5	4.2	4.4	5.4	5.8	5.4	6.5	1.3
Metaproterenol	9.9	10.6	9.1	7.6	8.9	8.7	9.7	7.9	8.3	9.2
Metoprolol	9.6	10.6	6.8	6.7	9.8	8.7	9.4	7.3	8.6	9.4
Miconazole	6.4	6.6	5.5	6.0	5.0	5.2	5.4	4.6	5.2	5.1
Morphine	8.2	10.1	5.7	5.5	7.5	7.6	5.6	6.0	4.3	4.9
Nafronyl	9.1	10.2	8.8	6.6	13.3	7.3	7.7	6.9	7.7	8.1
Nefopam	8.5	10.0	6.8	6.8	8.8	7.4	7.8	6.6	7.0	7.9
Niacine-	4.8	5.2	3.9	3.9	6.1	5.4	5.3	4.8	4.1	5.1
Nicotine	8.1	10.3	7.2	7.3	8.1	8.4	8.5	8.3	8.0	7.3
Nicotine+	3.2	5.2	1.7	1.8	2.4	1.6	1.7	1.4	1.3	-0.7
Nikethamide	3.5	5.2	2.0	2.4	3.6	2.5	2.6	2.0	1.8	2.1
Papaverine	6.4	6.0	3.9	4.4	4.7	4.9	5.0	3.5	4.0	7.3
p-Cl-amphetamine	9.9	10.4	9.0	9.0	10.0	9.2	8.7	8.8	8.2	8.9
Phenacaine	9.3	10.3	10.3	10.1	10.8	8.6	7.7	8.1	6.9	12.2
Phenylalanine-	8.9	9.5	9.7	9.3	10.1	9.4	9.2	8.4	8.1	9.3
Piroxicam	5.3	6.5	5.7	0.5	7.1	6.3	7.7	4.9	6.2	2.3
Prazosin	7.0	7.0	4.6	4.9	6.1	5.0	5.7	4.8	3.6	7.7
Procaine	9.1	10.2	8.6	6.7	10.9	8.6	8.3	8.5	9.1	9.2
Procaine+	2.0	4.6	-1.0	-0.7	-0.2	2.0	1.4	1.3	0.2	-1.9
Propanolol	9.6	10.6	5.2	6.8	8.8	8.3	8.4	7.5	7.7	8.5
Quinine	8.5	10.5	6.9	6.5	9.6	8.4	8.5	6.5	8.0	8.2
Sotalol	9.3	10.6	7.4	8.1	8.3	9.3	8.0	7.9	7.5	9.1
Sparteine	12.0	10.3	14.4	13.5	15.9	11.7	12.3	10.8	12.1	9.5
Tetracaine	8.5	10.2	9.3	7.8	10.7	9.6	9.0	9.1	9.3	8.7
Thenyldiamine	8.9	10.2	11.5	9.3	13.1	9.4	8.7	9.1	8.5	8.6
Trazodone	6.8	10.2	4.7	3.7	8.1	6.4	6.7	6.2	4.9	6.4
Trimipramine	9.4	10.2	11.9	10.2	13.7	10.2	10.2	9.4	10.1	8.1
Tryptophan-	9.1	9.5	9.6	9.2	11.3	9.5	9.8	8.9	8.4	10.6

Table 2: Root-mean-square-error (RMSE), statistical uncertainty (95% confidence limits in the RMSE, see SI for more information), and the maximum absolute error (Max AE) of the pKa (a) the pKa values in Table 1, (b) with cefradoxil removed, (c) with an empirical offset, and (d) using geometries optimized in the gas phase and zwittterions removed (Table S2). "COS" stands for COSMO.

	Ref pKa	PM6-DH+ COS	PM6 COS	PM7 COS	PM3 COS	AM1 COS	PM3 SMD	AM1 SMD	DFTB3 SMD
RMSE[a]	1.8	2.3	2.1	2.0	1.4	1.3	1.5	1.6	2.4
95% conf	1.4-2.1	1.8-2.7	1.6-2.4	1.6-2.4	1.1-1.6	1.1-1.6	1.2-1.8	1.3-1.9	1.9-2.8
Max AE	6.5	9.8	5.8	5.9	6.9	5.4	2.9	3.9	9.9
RMSE[b]	1.8	1.8	1.9	1.8	1.0	1.1	1.5	1.6	2.0
95% conf	1.4-2.1	1.4-2.2	1.5-2.3	1.4-2.2	0.8-1.2	0.9-1.3	1.2-1.8	1.2-1.9	1.6-2.3
Max AE	6.5	4.4	4.8	4.3	2.5	2.6	2.9	3.9	5.1
RMSE[c]	1.8	1.6	1.4	1.8	0.9	1.1	1.0	1.1	1.9
95% conf	1.4-2.1	1.3-1.9	1.1-1.7	1.4-2.1	0.7-1.1	0.8-1.2	0.8-1.2	0.9-1.3	1.5-2.2
Max AE	6.5	3.8	3.4	3.7	2.3	3.0	2.1	2.8	5.6
RMSE[d]	1.8	2.6	2.4	2.1	1.4	1.3	1.9	1.9	3.0
95% conf	1.4-2.1	2.0-3.0	1.8-2.8	1.6-2.5	1.1-1.7	1.0-1.5	1.5-2.2	1.5-2.2	2.4-3.6
Max AE	6.5	7.8	4.6	5.2	3.5	2.5	4.3	3.8	11.1

If the cefadroxil outlier is removed, the RMSE values for PM3/COSMO and AM1/COSMO drop to 1.0 ± 0.2 and 1.1 ± 0.2, while PM3/SMD and AM1/SMD remain at 1.5 ± 0.3 and 1.6 ± 0.3/0.4 pH units. Thus, without this outlier the COSMO-based predictions outperform the SMD-based predictions, as well as the null model. For pKa calculations where a zwitterionic state is not involved or proton transfer in a zwitterionic state is not observed then PM3/COSMO or AM1/COSMO is the best pKa prediction method, otherwise PM3/SMD or AM1/SMD should be used. The main reason for performing solution-phase geometry optimisations was the possible presence of zwitterions, so if a zwitterionic state is not involved then the geometry optimisations could potentially be done in the gas phase. Table 2 shows that PM3/COSMO and AM1/COSMO continue to perform best with RMSEs of 1.4 ± 0.3 and 1.3 ± 0.2/0.3 pH units, respectively (the pKa values can be found in Table S2). The largest difference in RMSEs is observed for PM3/COSMO(soln) and PM3/COSMO(gas) (0.4 pH units) and is larger than the composite error of 0.1 pH units for these two error. So using gas phase geometries for non-zwitterionic molecules leads to a statistically significant decrease in the accuracy of the pKa predictions.

Figure 2 shows that all semiempirical methods except PM7 tend to underestimate the pKa values. The mean signed errors for PM3/COSMO and AM1/COSMO are -0.4 and -0.5 pH units while they are -1.1 for both PM3/SMD and AM1/SMD (computed without cefadoxil). If these mean errors are included as an empirical correction to the pKa values then the accuracy of the COSMO- and SMD-based methods become statistically identical with RMSE values of between 0.9 and 1.1 pH units (Table 2). However, it remains to be seen whether these corrections are transferable to other sets of amines.

PM6-DH+-, PM6- and PM7-based methods

In addition to their chemical importance pKa values are also useful benchmarking tools that can help in identifying problems with theoretical methods. Here we compare the results for PM6-DH+/COSMO, PM6/COSMO- and PM7/COSMO-based methods to PM3/COSMO to gain some insight in to why these methods lead to less accurate pKa predictions with RMSE values of 1.9 compared to 1.0 (ignoring cefadroxil).

Compared to PM3, PM6-DH+ has two outliers: propranolol and lidocaine (Figure 2). For propanolol PM6-DH+ predicts a pKa value of 5.2, which is significantly lower than the experimental value of 9.6 and that predicted by PM3 (8.3). Comparison of the lowest free energy structures for the protonated state shown in Figure 3a-b shows that the PM6-DH+ structure is significantly more compact than the PM3 structure with the isopropylaminoethanol chain stacked on the face with the naphthalene group. This will lead to desolvation of the amine group and will lower the predicted pKa. This structure is also the lowest free energy structure for PM6 where the predicted pKa value is 6.8. So the compactness is not solely due to the dispersion interactions included in PM6-DH+, as one might expect, but these forces do contribute to the very low pKa value. It is important to emphasize that this does not necessarily imply that the dispersion interactions are overestimated by the DH+ corrected, but rather that they possibly are too large compared to the solute/solvent interactions in the COSMO solvation model when using PM6-DH+ to describe the solute. This general point also applies to the rest of the analyses presented below.

For lidocaine PM6-DH+ predicts a pKa value of 3.7 pH units, which is significantly lower than the experimental value of 7.9 and that predicted by PM3 (5.4). Comparison of the lowest free energy structures for the protonated state shown in Figure 3c-d shows that the NH-O hydrogen bond-like interaction observed in the PM3 structure is absent in the PM6-DH+ structure, which is consistent with a lower pKa value. The hydrogen bond,

which is also present in the lowest free energy PM6 structure, is replaced by non-polar interactions between methyl groups which presumably are stronger in PM6-DH+ due to the dispersion forces.

Compared to PM3, PM6 has one outliers (Figure 2), piroxicam, where PM6 predicts a pKa value of 0.5, which is significantly lower than the experimental value of 5.3 and that predicted by PM3 (6.3). Comparison of the lowest free energy structures for the protonated state shown in Figure 4 shows that the pyridine NH hydrogen bond to the amide O observed in the PM3 structure is replaced by a presumably unfavorable NH-HN interaction with the amide group, which indeed should lower the pKa considerably. Both PM3 and PM6 geometry optimisations are performed with exactly the same set of starting structures and it is not immediately clear why this arrangement leads to lowest free energy, but it is presumably due to an increase in the solvation energy.

Compared to PM3, PM7 has three outliers (Figure 2): spartein, trimipramine, and thenyldiamine. For propanolol PM7 predicts a pKa value of 15.9, which is significantly higher than the experimental value of 12.0 and that predicted by PM3 (11.7). Comparison of the lowest free energy structures for the protonated state shown in Figure 5a-b shows virtually no difference in structure. The same is found for the low free energy structures of the conjugate base and both protonation states of the reference molecule. The most likely explanation for the overestimation is therefore that the NH-N hydrogen bond strength is overestimated compared to PM3. This theory is further corroborated for trimiparine where PM7 predicts a pKa value of 13.7 pH units, which is significantly higher than the experimental value of 9.4 and that predicted by PM3 (10.2). Comparison of the lowest free energy structures for the protonated state shown in Figure 5c-d shows a NH-N hydrogen bond for the PM7 structure, which is absent in the PM3 structure. This structural difference is consistent with both the higher pKa and an overestimation of NH-N hydrogen bond strength by

PM7. Finally, for thenyldiamine PM7 predicts a pKa value of 13.1 pH units, which again is significantly higher than the experimental value of 8.9 and that predicted by PM3 (9.4). The main difference in structure between the free energy minima (Figure 5e-f) is an apparently stronger interaction between the thiophene ring and the amine in the PM7 structure, which, if anything, should desolvate the amine group and lower the pKa value. The most likely explanation for the overestimation is thus an overestimation of the NH-N hydrogen bond as in the the other two cases.

DFTB3/SMD

Compared to PM3/COSMO, DFTB3/SMD has five outliers (Figure 2) and here we focus on the two with the largest errors: guanethidine and mechlorethamine. For guanethidine DFTB3 predicts a pKa value of 16.2 pH units, which is significantly higher than the experimental value of 11.4 and that predicted by PM3 (13.2). Comparison of the lowest free energy structures for the protonated state shown in Figure 6a-b shows that they are quite similar with a NH-N hydrogen bond, but with the 7-membered ring in a slightly different conformation. The hydrogen bond length in the DFTB3 structure is 2.33 Å, which is significantly shorter than the 2.56 Åin the PM3 structure. A stronger hydrogen bond is consistent with a higher pKa, but the errors for DFTB3 are not unusually larger for, for example, sparteine, trimipramine, and thenyldiamine. One possibility is that it is only guanine NH hydrogen bond strengths that are overestimated but this can not be verified with the current set of molecules.

For mechlorethamine DFTB3 predicts a pKa value of 1.3 pH units, which is significantly lower than the experimental value of 6.4 and that predicted by PM3 (5.4). Comparison of the lowest free energy structures for the protonated state shown in Figure 6c-d shows overall similar structures. In both cases the amine hydrogen is surrounded by the two chlorine atoms, which lowers the pKa value due to desolvation. However, closer inspection of the

structures reveal that for the DFTB3 structure the chlorine atoms are significantly closer together and one of the chlorine atoms is significantly closer to the amine hydrogen. These structural differences are consistent with greater desolvation in the DFTB3 structure and, hence, a lower pKa value.

With regard to DFTB3 it is also noteworthy that two molecules fragment in the DFTB3 gas phase geometry optimisations: in the case of the niacin zwitterion CO_2 is eliminated while for the protonated form of sotalol CH_3SO_2 is eliminated. Barrier-less CO_2 has been previously observed for DFTB3 for model systems of L-aspartate α-decarboxylase[42] and is presumably due to the 16.8 kcal/mol error in the atomisation energy of CO_2 for DFTB3[10].

Prediction of dominant protonation state

One of the main uses of pKa values is the prediction of the correct protonation state at physiological pH (7.4), i.e. determining whether the predicted pKa value is above or below 7.4. Here (ignoring cefadroxil) PM3/COSMO performs best by getting it right 94% of the time, compared to 90%, 79%, and 92% for AM1/COSMO, PM3/SMD, and the null model. Thus, only PM3/COSMO outperforms the null model. PM3/COSMO fails in three cases, labetalol, lidocaine, and nafronyl, where PM3/COSMO predicts pKa values of 7.5, 5.4, and 7.3, respectively and the corresponding experimental values are 7.3, 7.9, and 9.1 pH units. The null model fails in four cases, clozapine (amide nitrogen), labetalol, mechlorethamine, and trazodone, where the null model predicts pKa values of 10.3, 10.6, 10.0, and 10.2 and the corresponding experimental values are 3.9, 7.3, 6.4, and 6.8 pH units, respectively. Thus, both methods fail for only one ionizable site where the experimentally measured pKa value is significantly different from physiological pH.

Table 3: Root-mean-square-error (RMSE), statistical uncertainty (95% confidence limits) in the RMSE, and the maximum absolute error (Max AE) of the pKa for the pKa values listed in Table S3

	PM3 COS	PM3 COS*	DFT	Chem Axon	Epik	ACD
RMSE	1.0	0.9	0.7	0.7	0.7	0.6
95% conf	0.8-1.2	0.7-1.1	0.5-0.8	0.6-0.9	0.6-0.8	0.5-0.7
Max AE	2.5	2.3	1.9	2.8	3.6	2.0

Comparison to other pKa prediction methods

Figure 7 and Table 3 compares the two best semiempirical methods PM3/COSMO and PM3/COSMO*, where the pKa values are shifted to make the average error zero, to the DFT results of Eckert and Klamt[6] and three popular empirical pKa prediction methods. In all cases cefadroxil has been removed. The RMSE values of the DFT and empirical methods are 0.6-0.7, 0.2-0.3 pH units lower than the best semiempirical method PM3/COSMO*. Thus, the accuracy of DFT results are statistically equivalent to the empirical methods, while the semiempirical methods are statistically worse. The good performance of the empirical methods for this set of molecules is not surprising. The set represents well known and prototypical drug molecules whose pKa values have been known for a long time and many of the molecules are likely included in the parameterization of the empirical methods. For example, many of the molecules are taken from the set collected by Klici et al.[7], which is also included in the training set used to develop Epik.[34] It is therefore gratifying to see that the DFT results by Eckert and Klamt[6], which only contains two adjustable parameters determined using the different set of data, are just as accurate albeit at a much higher computational cost. The computational cost of the DFT method is ca 1000 times larger than that of the semiempirical methods, while the computational cost of the empirical methods is essentially zero compared to the semiempirical methods.

As mentioned in the introduction, one potential use of the QM-based pKa prediction methods is for cases where the empirical methods fail. Figure 7 shows that, for example,

ChemAxon has two outliers while Epik has one outlier where the error is larger than for the QM-based predictions. The Epik outlier is observed for sparteine, which is also the one of the outliers observed for ChemAxon, as well as one of the largest errors observed for ACD. The absolute errors for these methods range from 1.9 to 3.6, while the errors for PM3/COSMO(*) and DFT are between 0.2 and 0.4 pKa units. Similarly, the other ChemAxon outlier is observed for labetalol, which is also gives rise to the second and third largest error for Epik and ACD, respectively. The absolute errors for these methods range from 1.6 to 2.5, while the errors for PM3/COSMO(*) and DFT are between 0.0 and 0.7 pKa units. These cases suggest that QM-based pKa prediction methods can be of practical use despite their comparatively high computational cost.

Timings

A MOPAC-based geometry optimization requires no more than about 10-20 CPU seconds on a single core CPU even for the largest molecules considered here (e.g. clozapine), whereas corresponding GAMESS optimizations take about 60-90 seconds. Thus, using 20 different starting geometries for each protonation state a pKa value can be predicted in a few CPU minutes using a single 12-CPU node. In practice the wall-clock time is longer due to the overhead involved in having all cores write output files to disk simultaneously. Similarly, most queuing software has an some computational overhead which becomes noticeable when a large number of sub-minute jobs are submitted simultaneously. These general problems need to be addressed if semiempirical methods are to be used efficiently in very large-scale high throughput studies. Nevertheless, fast pKa predictions for 100-1000s of molecules is feasible with the current setup and relatively modest computational resources.

Summary and Outlook

This study uses PM6-DH+/COSMO, PM6/COSMO, PM7/COSMO, PM3/COSMO, AM1/COSMO, PM3/SMD, AM1/SMD, and DFTB3/SMD to predict the pKa values of 53 amine groups in 48 drug-like compounds. The approach uses isodesmic reactions where the pKa values is computed relative to a chemically related reference compound for which the pKa value has been measured experimentally or estimated using an standard empirical approach. Both gas phase and solution phase geometry optimisations are tested. The AM1- and PM3-based methods using solution phase geometries perform best with RMSE values of 1.4 - 1.6 pH units that have uncertainties of 0.2-0.3 pH units, which make them statistically equivalent. However, for all but PM3/SMD and AM1/SMD the RMSEs is dominated by a single outlier, cefadoxil, caused by proton transfer in the zwitterionic protonation state. If this outlier is removed, the RMSE values for PM3/COSMO and AM1/COSMO drop to 1.0 $\pm$ 0.2 and 1.1 $\pm$ 0.3, while PM3/SMD and AM1/SMD remain at 1.5 $\pm$ 0.3 and 1.6 $\pm$ 0.3/0.4 pH units. Thus, without this outlier the COSMO-based predictions outperform the SMD-based predictions, so for pKa calculations where a zwitterionic state is not involved or proton transfer in a zwitterionic state is not observed then PM3/COSMO or AM1/COSMO is the best pKa prediction method, otherwise PM3/SMD or AM1/SMD should be used. Thus, fast and relatively accurate pKa predictions for 100-1000s of molecules is feasible with the current setup and relatively modest computational resources.

For the current study the reference molecules were selected by hand to match the local structure around the ionizable as much as possible for most molecules to maximize the cancellation of error and improve accuracy as much as possible. This approach will work well when the pKa of a small number of molecules is needed or if the effect of substituents on the pKa of an ionizable group in a target molecule is to be investigated. However, for high throughput pKa prediction for a very large and diverse set of molecules it will not always be practically possible to identify closely related reference molecules and for such a case the

overall accuracy is likely to be worse than reported here. How much worse remains to be seen but recalculating the PM3/COSMO pKa values (without cefadroxil) using only nine reference compounds (Table S4) results in an RMSE value of 1.2, i.e. only 0.2 pH units higher than that computed using 26 reference values - an encouraging result. After this paper was submitted, Bochevarov et al.[8] published a paper on DFT-based pKa prediction where they defined linear regression parameters for roughly 100 different ionizable functional groups and outlined a hierarchical model for choosing the most appropriate parameter set. This interesting approach could serve as a basis for a defining a more generally applicable set of reference molecules in future work. The current implementation also relies on manual selection of the protonation state of other ionizable groups, which in cases like cimetidine requires expert knowledge. In the general case this step needs to be automated by generating all possible protonation isomers for a given protonation state and selecting the one with the lowest free energy. Work on full automation of the process is ongoing.

Acknowledgement

JHJ thanks Anders Christensen for help with the Seaborn plotting package, Jimmy Kromann for help with submit scripts, and Anthony Nicholls for helpful answers to questions about RMSE uncertainties and statistical significance. JHJ also thanks Mark S. Gordon for reasons too numerous to list here.

Supporting Information Available

All input and output files and python scripts that automate most parts of the computations

This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Settimo, L.; Bellman, K.; Knegtel, R. M. A. Comparison of the Accuracy of Experimental and Predicted pKa Values of Basic and Acidic Compounds. *Pharm Res* **2013**, *31*, 1082–1095.

(2) Rupakheti, C.; Al-Saadon, R.; Zhang, Y.; Virshup, A. M.; Zhang, P.; Yang, W.; Beratan, D. N. Diverse Optimal Molecular Libraries for Organic Light-Emitting Diodes. *J. Chem. Theory Comput.* **2016**, *12*, 1942–1952.

(3) Gomez-Bombarelli, R.; Duvenaud, D.; Hernandez-Lobato, J. M.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *arXiv* **2016**, *arXiv:1610.02415*.

(4) Husch, T.; Korth, M. Charting the known chemical space for non-aqueous lithium–air battery electrolyte solvents. *Phys. Chem. Chem. Phys.* **2015**, *17*, 22596–22603.

(5) Ho, J. Predicting pKa in Implicit Solvents: Current Status and Future Directions. *Aust. J. Chem.* **2014**, *67*, 1441.

(6) Eckert, F.; Klamt, A. Accurate prediction of basicity in aqueous solution with COSMO-RS. *J. Comput. Chem.* **2005**, *27*, 11–19.

(7) Klici, J. J.; Friesner, R. A.; Liu, S.-Y.; Guida, W. C. Accurate Prediction of Acidity Constants in Aqueous Solution via Density Functional Theory and Self-Consistent Reaction Field Methods. *The Journal of Physical Chemistry A* **2002**, *106*, 1327–1335.

(8) Bochevarov, A. D.; Watson, M. A.; Greenwood, J. R.; Philipp, D. M. Multiconformation, Density Functional Theory-Based pKa Prediction in Application to Large, Flexible Organic Molecules with Diverse Functional Groups. *Journal of Chemical Theory and Computation* **2016**, *12*, 6001–6019.

(9) Stewart, J. J. P. Application of the PM6 method to modeling proteins. *Journal of Molecular Modeling* **2008**, *15*, 765–805.

(10) Gaus, M.; Cui, Q.; Elstner, M. DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB). *J. Chem. Theory Comput.* **2011**, *7*, 931–948.

(11) Kromann, J. C.; Larsen, F.; Moustafa, H.; Jensen, J. H. Prediction of pKa values using the PM6 semiempirical method. *PeerJ* **2016**, *4*, e2335.

(12) Korth, M. Third-Generation Hydrogen-Bonding Corrections for Semiempirical QM Methods and Force Fields. *J. Chem. Theory Comput.* **2010**, *6*, 3808–3816.

(13) Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *Journal of Molecular Modeling* **2007**, *13*, 1173–1213.

(14) Stewart, J. J. P. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *Journal of Molecular Modeling* **2012**, *19*, 1–32.

(15) Stewart, J. J. P. Optimization of parameters for semiempirical methods I. Method. *J. Comput. Chem.* **1989**, *10*, 209–220.

(16) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *Journal of the American Chemical Society* **1985**, *107*, 3902–3909.

(17) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk

Dielectric Constant and Atomic Surface Tensions. *The Journal of Physical Chemistry B* **2009**, *113*, 6378–6396.

(18) Klamt, A.; Schrmann, G. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *Journal of the Chemical Society, Perkin Transactions 2* **1993**, 799–805.

(19) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. et al. General atomic and molecular electronic structure system. *Journal of Computational Chemistry* **1993**, *14*, 1347–1363.

(20) Steinmann, C.; Bldel, K. L.; Christensen, A. S.; Jensen, J. H. Interface of the polarizable continuum model of solvation with semi-empirical methods in the GAMESS program. *PloS one* **2013**, *8*, e67725.

(21) Nishimoto, Y. DFTB/PCM Applied to Ground and Excited State Potential Energy Surfaces. *J. Phys. Chem. A* **2016**, *120*, 771–784.

(22) Gaus, M.; Lu, X.; Elstner, M.; Cui, Q. Parameterization of DFTB3/3OB for Sulfur and Phosphorus for Chemical and Biological Applications. *J. Chem. Theory Comput.* **2014**, *10*, 1518–1537.

(23) Lu, X.; Gaus, M.; Elstner, M.; Cui, Q. Parametrization of DFTB3/3OB for Magnesium and Zinc for Chemical and Biological Applications. *The Journal of Physical Chemistry B* **2015**, *119*, 1062–1082.

(24) Kubillus, M.; Kubař, T.; Gaus, M.; Řezáč, J.; Elstner, M. Parameterization of the DFTB3 Method for Br Ca, Cl, F, I, K, and Na in Organic and Biological Systems. *J. Chem. Theory Comput.* **2015**, *11*, 332–342.

(25) Baker, J.; Kessi, A.; Delley, B. The generation and use of delocalized internal coordinates in geometry optimization. *The Journal of Chemical Physics* **1996**, *105*, 192.

(26) ACE/JChem, ACE and JChem acidity and basicity calculator. **2016**, *https://epoch.uky.edu/ace/public/pKa.jsp*.

(27) Hirt, R. C.; Schmitt, R. G.; Strauss, H. A.; Koren, J. G. Spectrophotometrically Determined Ionization Constants of Derivatives of Symmetric Triazine. *Journal of Chemical & Engineering Data* **1961**, *6*, 610–612.

(28) Charton, M. The Application of the Hammett Equation to Amidines. *The Journal of Organic Chemistry* **1965**, *30*, 969–973.

(29) Baba, T.; Matsui, T.; Kamiya, K.; Nakano, M.; Shigeta, Y. A density functional study on the pKaof small polyprotic molecules. *International Journal of Quantum Chemistry* **2014**, *114*, 1128–1134.

(30) Lordi, N. G.; Christian, J. E. Physical Properties and Pharmacological Activity: Antihistaminics. *Journal of the American Pharmaceutical Association (Scientific ed.)* **1956**, *45*, 300–305.

(31) Prankerd, R. J. *Profiles of Drug Substances, Excipients and Related Methodology*; Elsevier BV, 2007; pp 1–33.

(32) Niazi, M. S. K.; Mollin, J. Dissociation Constants of Some Amino Acid and Pyridinecarboxylic Acids in Ethanol–H 2 O Mixtures. *Bulletin of the Chemical Society of Japan* **1987**, *60*, 2605–2610.

(33) Landrum, G. RDKit. **2016**, *http://rdkit.org/*.

(34) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: a software program for pK a prediction and protonation state generation for drug-like molecules. *Journal of Computer-Aided Molecular Design* **2007**, *21*, 681–691.

(35) Greenwood, J. R.; Calkins, D.; Sullivan, A. P.; Shelley, J. C. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *Journal of Computer-Aided Molecular Design* **2010**, *24*, 591–604.

(36) Nicholls, A. Confidence limits, error bars and method comparison in molecular modeling. Part 1: The calculation of confidence intervals. *Journal of Computer-Aided Molecular Design* **2014**, *28*, 887–918.

(37) Nicholls, A. Confidence limits, error bars and method comparison in molecular modeling. Part 2: comparing methods. *Journal of Computer-Aided Molecular Design* **2016**, *30*, 103–126.

(38) Jensen, J. H. Which method is more accurate? or errors have error bars. *PeerJ Preprints* **2017**, *5*, e2693v1.

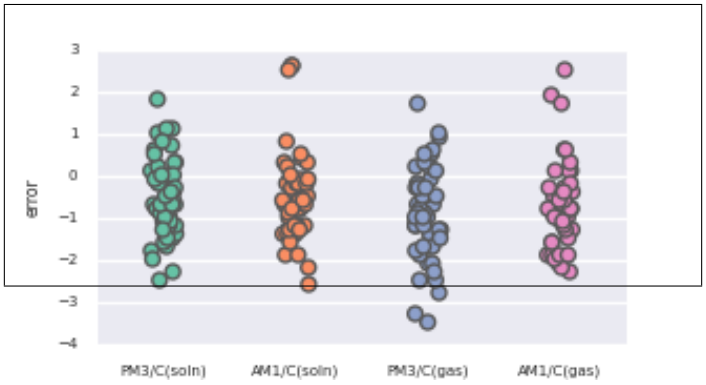
(39) Wang, W.; Pu, X.; Zheng, W.; Wong, N.-B.; Tian, A. Some theoretical observations on the 1:1 glycine zwitterion–water complex. *Journal of Molecular Structure: THEOCHEM* **2003**, *626*, 127–132.

(40) Bachrach, S. M. Microsolvation of Glycine: A DFT Study. *J. Phys. Chem. A* **2008**, *112*, 3722–3730.

(41) Kayi, H.; Kaiser, R. I.; Head, J. D. A theoretical investigation of the relative stability of hydrated glycine and methylcarbamic acid—from water clusters to interstellar ices. *Physical Chemistry Chemical Physics* **2012**, *14*, 4942.

(42) Kromann, J. C.; Christensen, A. S.; Cui, Q.; Jensen, J. H. Towards a barrier height benchmark set for biologically relevant systems. *PeerJ* **2016**, *4*, e1994.

Graphical TOC Entry



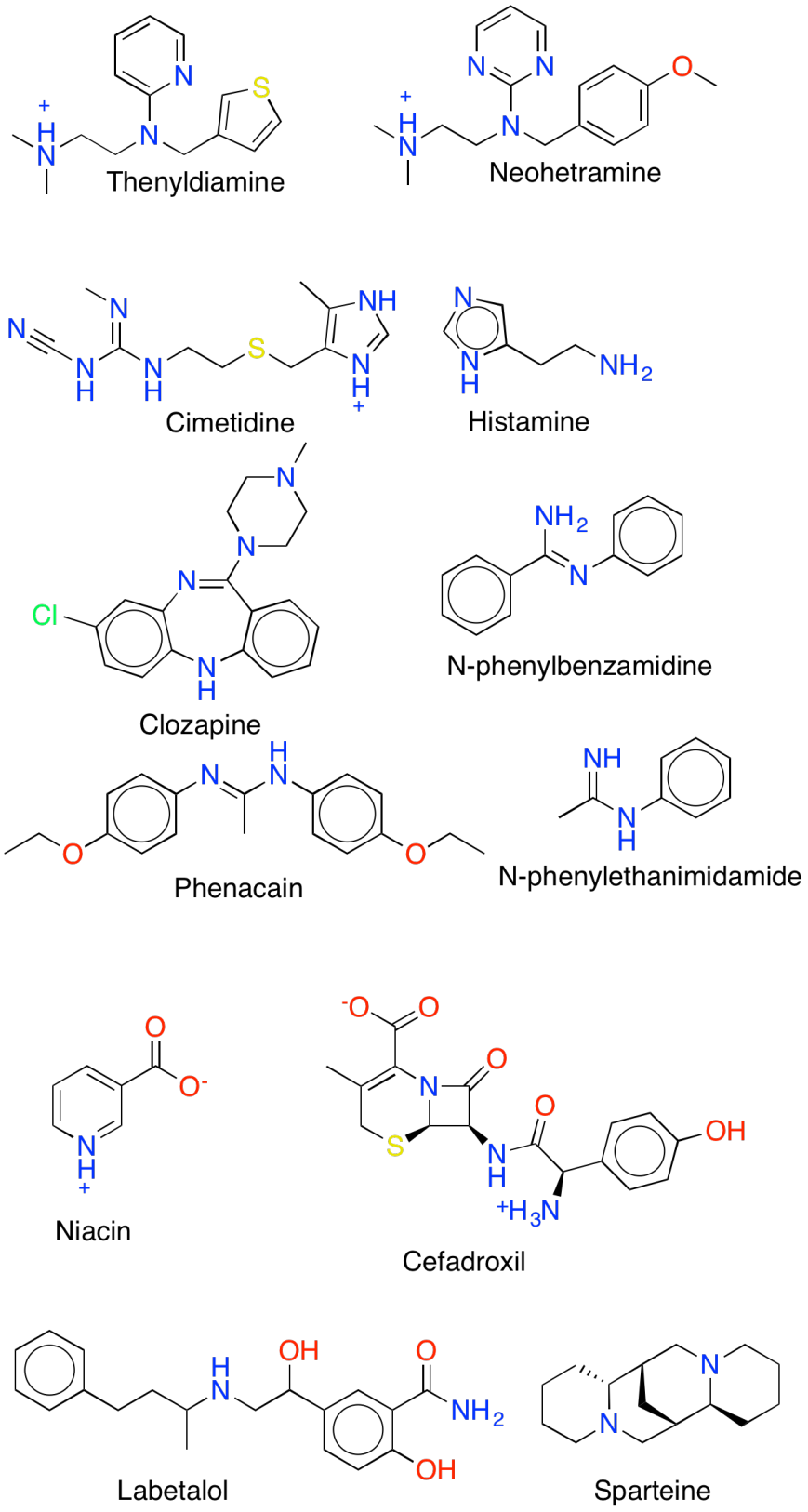


Figure 1: Some of the molecules referred to in the text

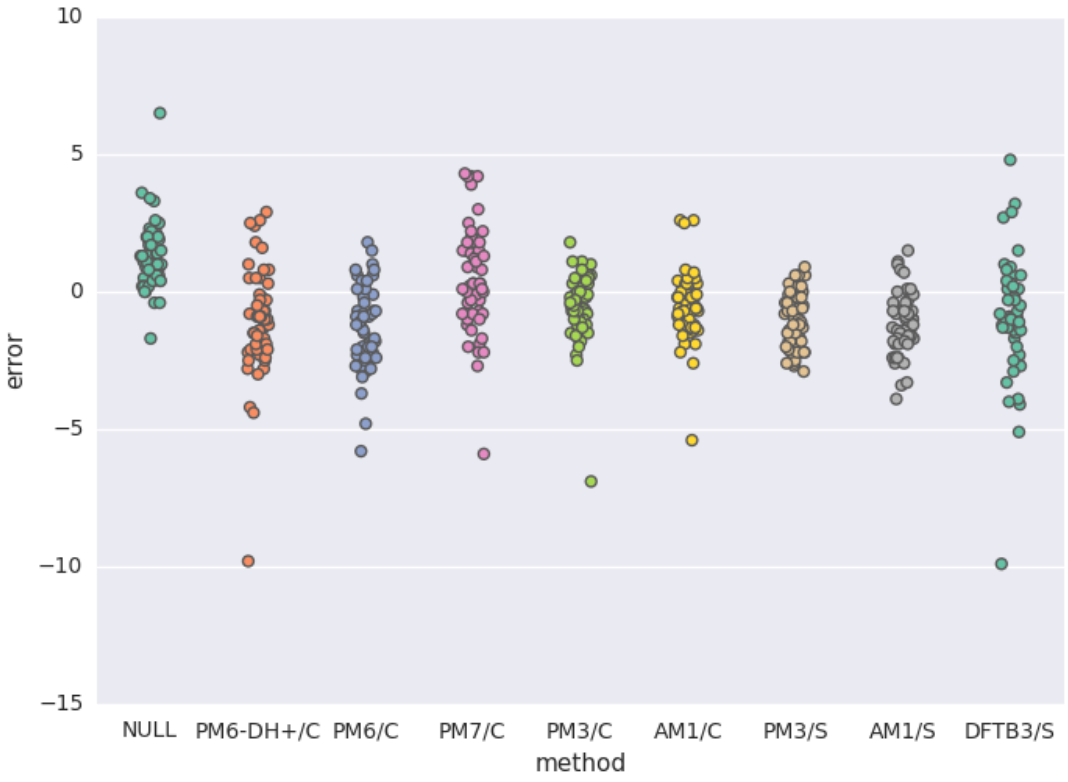


Figure 2: Plot of the errors of the predicted pK_a values ($pK_a - pK_a^{Exp}$). "C" and "S" stand for COSMO and SMD, respectively.

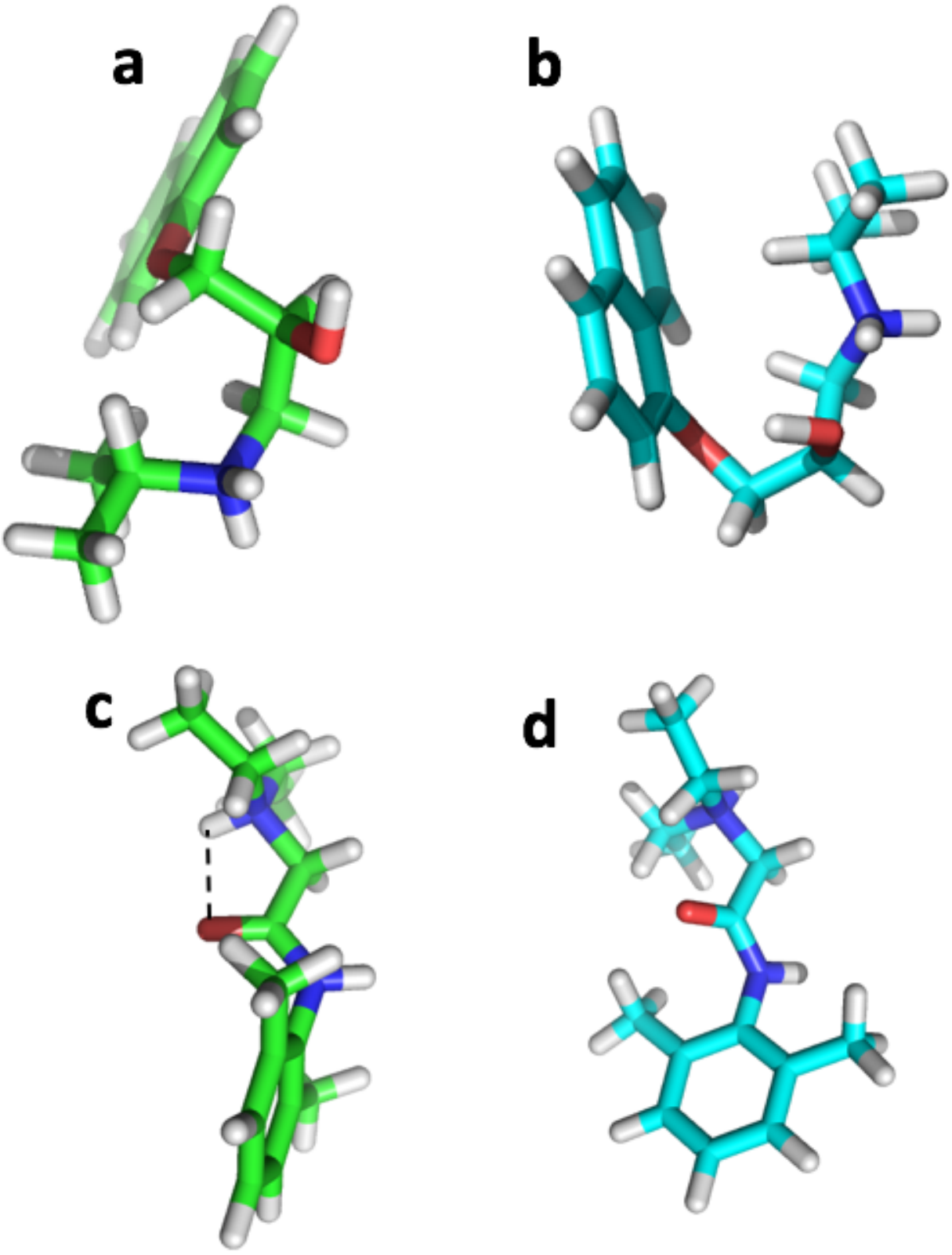


Figure 3: Lowest free energy conformations of (a-b) propanolol and (c-d) lidocaine at the PM3/COSMO (a and c) and PM6-DH+/COSMO (b and d) level of theory. Hydrogen bonds are indicated with dashed lines.

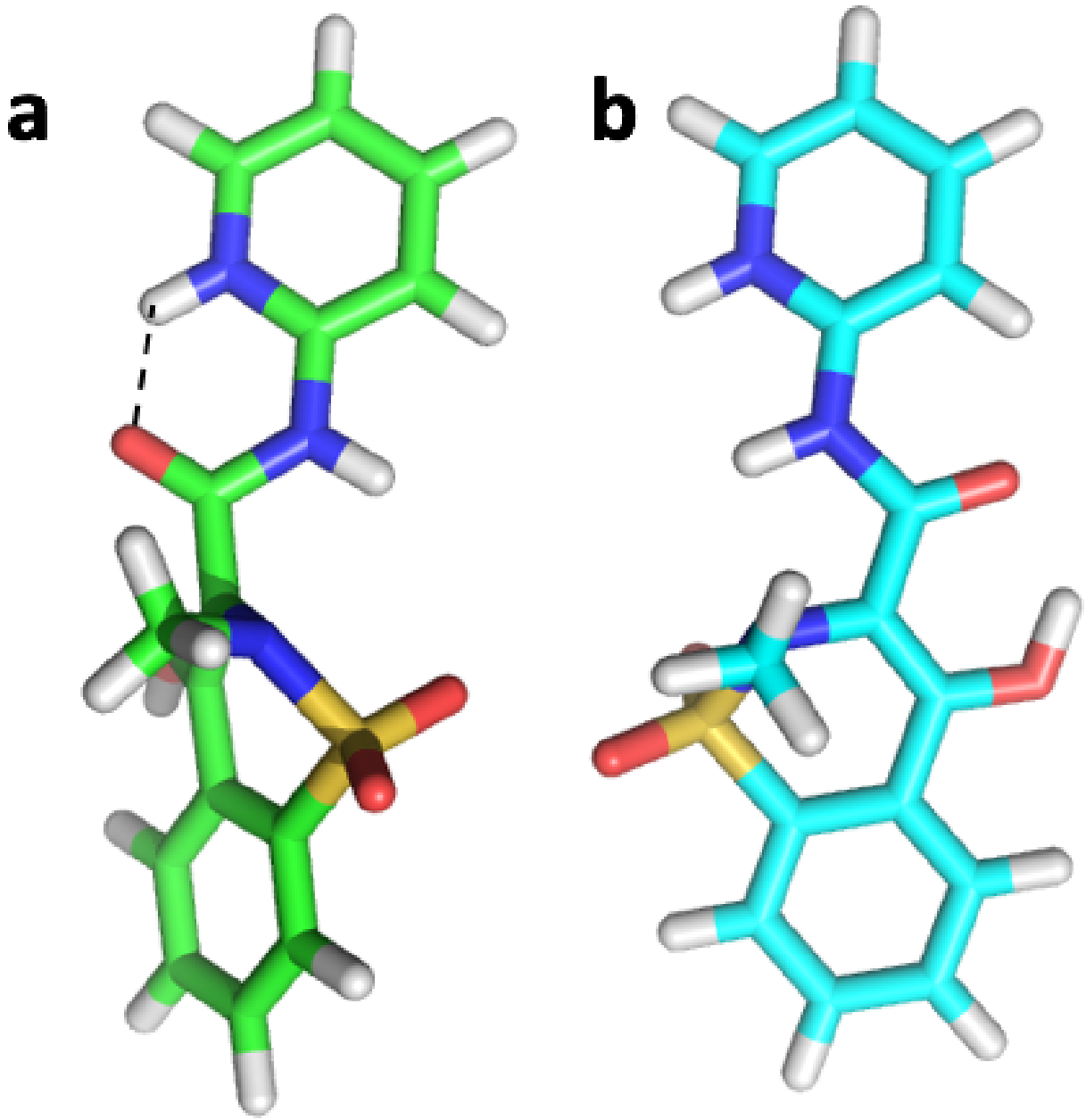


Figure 4: Lowest free energy conformations of piroxicam at the (a) PM3/COSMO and (b) PM6/COSMO level of theory. Hydrogen bonds are indicated with dashed lines.

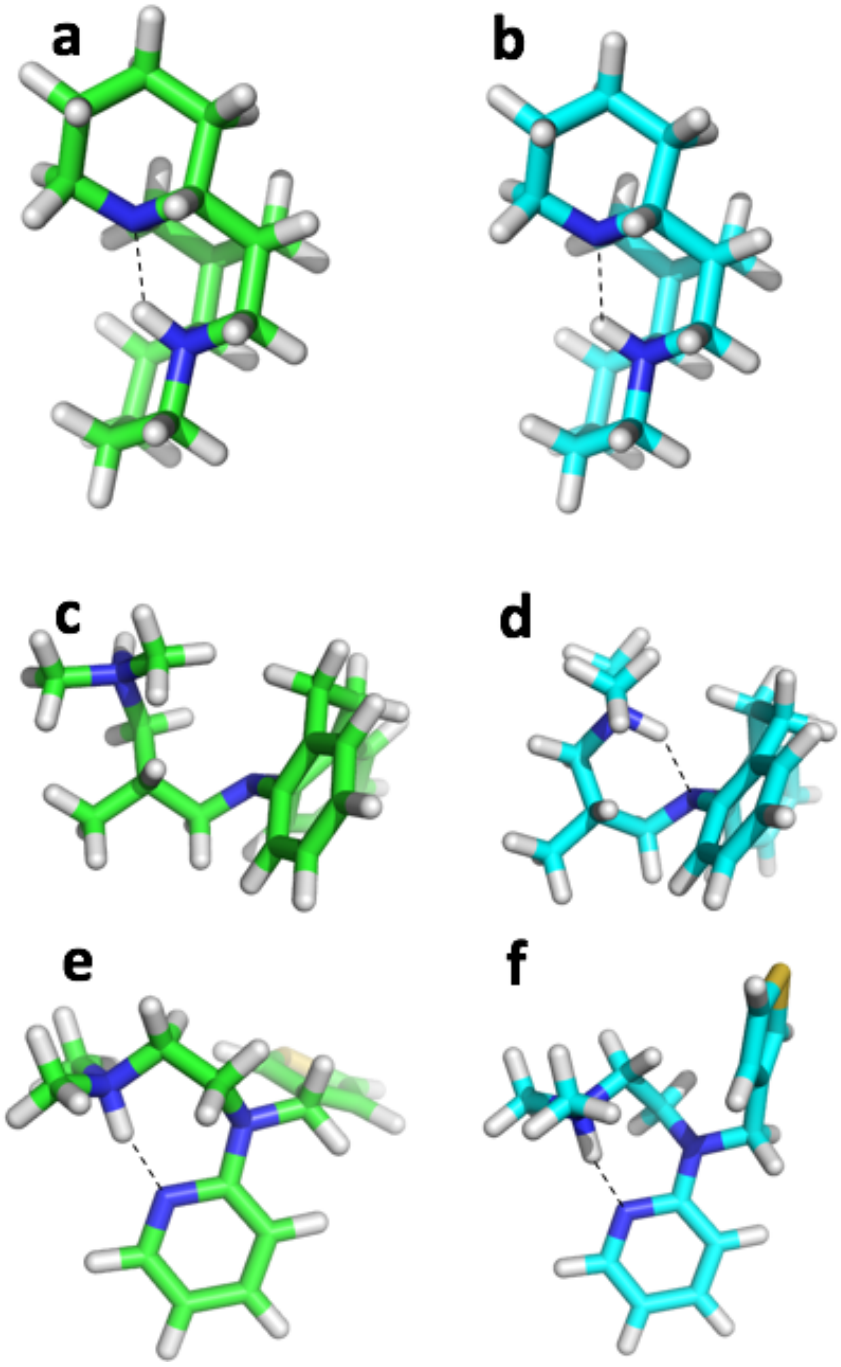


Figure 5: Lowest free energy conformations of (a-b) sparteine, (c-d) trimipramine, and (e-f) thenyldiamine at the PM3/COSMO (a, c, and e) and PM7/COSMO (b, d, and f) level of theory. Hydrogen bonds are indicated with dashed lines.

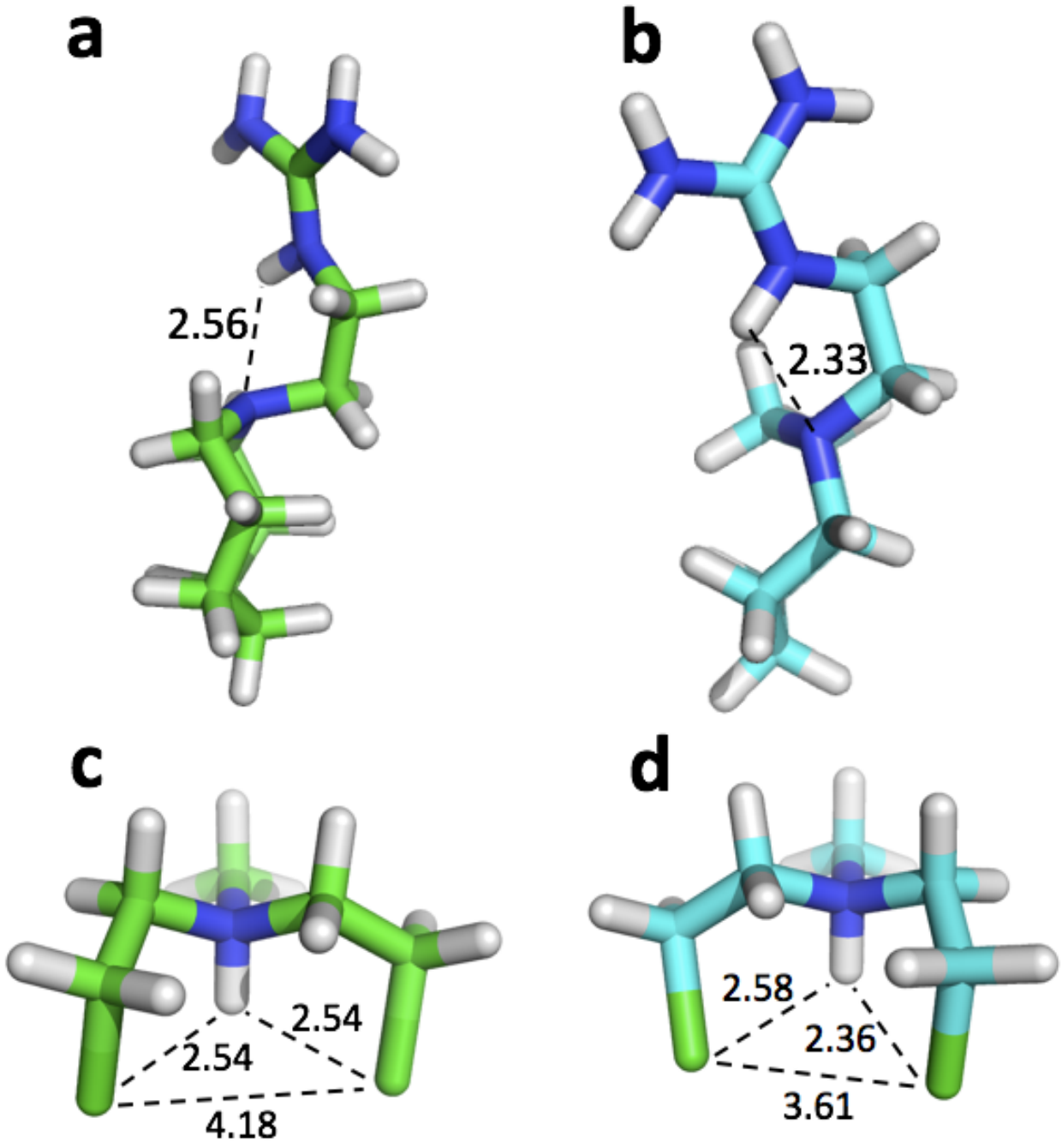


Figure 6: Lowest free energy conformations of (a-b) guanethidine and (c-d) mechlorethamine at the PM3/COSMO (a and c) and DFTB3/SMD (b and d) level of theory. Distances are given in Å.

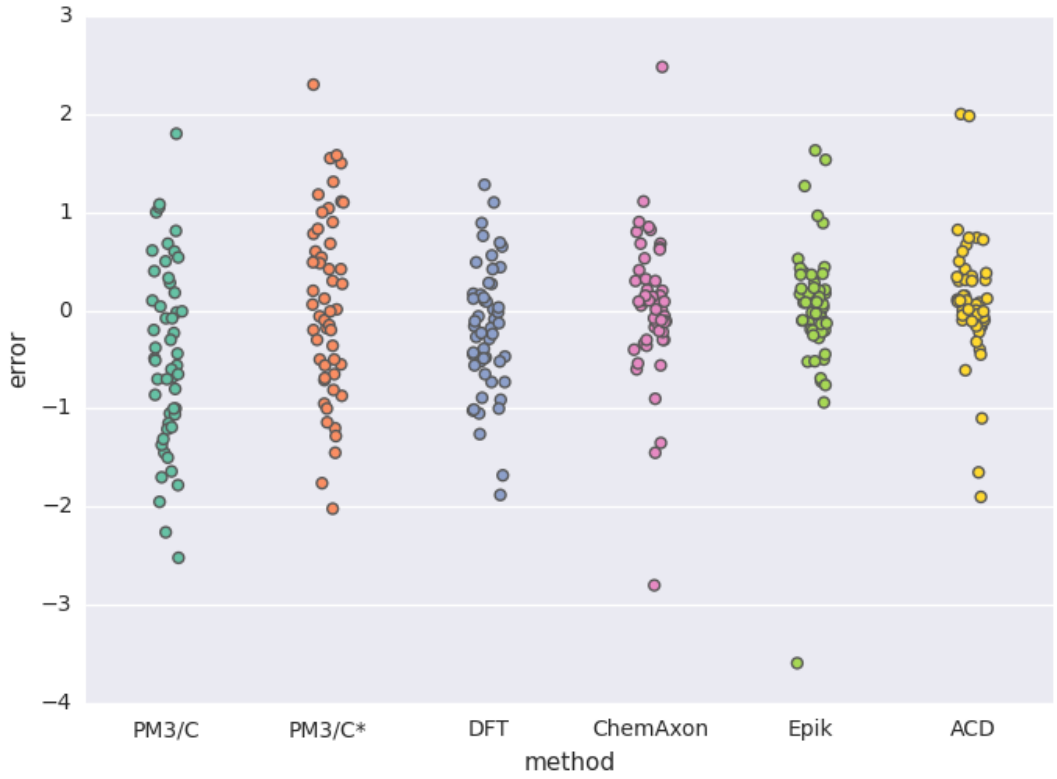


Figure 7: Plot of the errors of the predicted pK_a values ($pK_a - pK_a^{Exp}$). "C" stands for COSMO and "*" indicates that the pKa values have been shifted to make the average error zero